

OCR Software

We may be a long way from the paperless office, but a decent OCR package can help you to digitise at least some of your paperwork. Ben Pitt examines the options

Computers may eventually relegate paper to history, but it won't happen any time soon. Books, CVs, bills, receipts – even computing magazines – still appear in print, most of the time at least. However, you can't do much with printed text other than read it. Digital text, however, is the epitome of malleability. You can edit it, reformat it, translate it into another language, email it across the globe and print out a hundred copies. Best of all, you can search it. Imagine being able to scan through every word on every bit of paper in your home or office in less than a second. The software we've tested this month doesn't just make it possible; it makes it easy.

Optical character recognition (OCR) software turns scans into editable, searchable text. The four packages here are mature applications that have little difficulty deciphering text from a letter or magazine page. We gave them some tougher challenges to separate the great from the good: converting very small text, unusual fonts, light-coloured text on white and black backgrounds and pages comprising complex layouts of words and images. We rated software according to how accurate its results were first time, but also looked for how quick and easy it was to identify and correct errors.

Modern OCR software does more than just turn scans into text. All four packages can also preserve images and combine them with the converted text, complete with formatting, to create an accurate record of the original document. Output options vary, but all four packages are comprehensive, with support for Microsoft Word and Excel, Adobe Acrobat (PDF), Rich Text Format (RTF) and, in most cases, more besides.

One of the most useful techniques for archiving paperwork is to save a PDF of the scanned bitmap with the recognised text hidden underneath. This means text can be searched and copied to other applications, but you also have a permanent record of the original document as an image file. Therefore, you don't need to worry about the OCR being 100 per cent accurate. Happily, all four packages can produce this type of file.

Don't be put off by packages that don't include search facilities. Readiris's search utility is available for free as Copernic Desktop Search (www.copernic.com) and works equally well with any OCR package. Alternatively, try Google Desktop Search. Both of these utilities keep an index of the contents of your files, including text within documents. This means that results appear in a fraction of a second.

ABBYY FineReader 8.0 Professional Edition



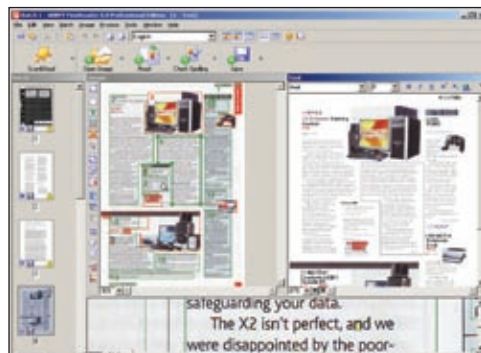
£71 inc VAT

FineReader is a no-frills OCR application, its only extra feature being a screen-capture utility. We can't think of many times that this utility would be useful, but capturing text from Flash websites is one exception.

The main application looks neat and friendly, with four buttons for importing, reading, checking spelling and saving, plus a fifth for picking automated routines. It's easy to create routines with the Automation Wizard, but there are a few inconveniences, such as scans being handled by the scanner's TWAIN driver rather than FineReader itself, and it can't generate unique filenames automatically. As a result, one-click automated processing isn't possible, but the software is still quick and easy to use. Export options are comprehensive, and FineReader was the only software we tested to produce Word documents with the Language set to English (UK) rather than English (US).

Character recognition was slow in our small text

test and worse for magazine pages. However, accuracy was often the best of the packages here. There was perfect accuracy in most tests and 92 per cent accuracy from 4-point text scanned at 300dpi. The other packages managed between 79 and five per cent accuracy in the same test. The



▲ Abbyy's FineReader 8.0 is neat and friendly, with accurate character recognition and formatting

spellchecker made it easy to make amendments to produce a 100 per cent accurate file. You can edit text freely in the editor window, but only uncertain characters and not unknown spellings are highlighted.

Formatting was preserved well, coping perfectly with complex layouts such as text wrapped around pictures. Fonts were often preserved, too, although condensed fonts were reproduced by bunching normal-width letters together, making them hard to read. Worse, the software was overzealous in its attempts to convert text, sometimes detecting text-like shapes that were

part of an image. However, it wasn't hard to fix this by defining text and image zones manually.

FineReader is a little slow at the recognition and spellchecking stages. Still, the accuracy of the character recognition and formatting means it's quick at producing usable results.

SUMMARY

- ✓ Highly accurate recognition
- ✓ Excellent formatting preservation
- ✗ Relatively slow

OCR SOFTWARE Requires Windows 2000/XP/2003, 500MHz processor, 128MB RAM, 350MB disk space
PRODUCT CODE ABB032
SUPPLIER www.pcwb.com DETAILS www.abbyy.com

THIS MONTH

ABBYY FineReader 8.0 Professional Edition

Page 118

I.R.I.S. Readiris Pro 11

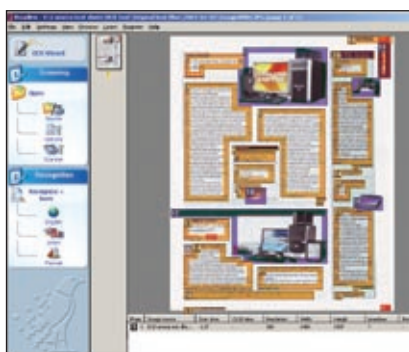
Page 119

Nuance Scansoft OmniPage 15

Page 119

Nuance Scansoft PaperPort 11

Page 119



▲ Readiris Pro 11's most useful extra is Desktop Search, the ideal accompaniment to OCR

I.R.I.S. Readiris Pro 11

★★★

£52 inc VAT

Readiris Pro looks friendly and approachable but its interface has room for improvement. Correcting errors is awkward, with just a dialog box and no free editing available. This is particularly frustrating when you can see mistakes in a word but are permitted to change only the highlighted character. Because spellchecking and saving are combined into a single action, you must perform the former again if you save in a different format.

Character-recognition accuracy was rarely less than 100 per cent for letters and magazine pages, but small and unusual fonts caused serious problems, often producing text with so many mistakes that it wasn't worth fixing. White text on a black background went completely unnoticed. Formatting was generally well preserved, but the software had very strange ideas about how text flowed around images, on one occasion jumping from the first to third to second paragraphs.

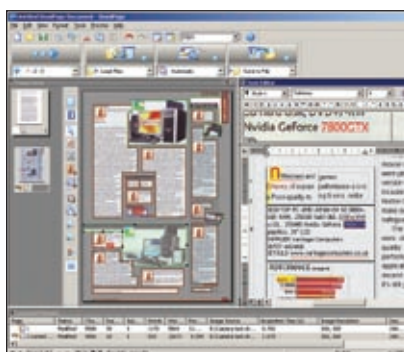
The package also includes Cardiris LE for extracting information from business cards, although this Limited Edition supports up to only 10 cards so it's really a demo for the £60 full version. PhotoScore MIDI Lite turns scans of sheet music into MIDI files and works well, but few people will use this feature.

I.R.I.S. Desktop Search is the most useful extra feature. It catalogues files on your hard disk, including the contents of text-based documents, for extremely fast searching. It's delightfully easy to use and is the ideal accompaniment to an OCR application. However, you don't need to buy Readiris Pro to get it, as it's a rebranded version of Copernic Desktop Search, which is available for free from www.copernic.com. The poor OCR accuracy and error correction make it impossible to recommend Readiris Pro II.

SUMMARY

- ✓ Excellent search facilities
- ✗ Relatively poor OCR accuracy
- ✗ Correcting errors is awkward

OCR SOFTWARE Requires Windows 98 SE/Me/2000/XP, 486 processor, 64MB RAM (128MB recommended), 105MB disk space
PRODUCT CODE SRISTPAPCUK110
SUPPLIER www.morecomputers.co.uk
DETAILS www.irislink.com



▲ Nuance OmniPage 15's Text Editor window helpfully underlines words not in its dictionary

NUANCE Scansoft OmniPage 15

★★★★★

£67 inc VAT



OmniPage is available in two versions: Professional, which costs around £320 including VAT, and the standard version reviewed here. It's simple to figure out, with three buttons for import, recognition and export, and a fourth to automate the other three. Creating automated routines from scratch is easy, but editing existing ones is less so.

We like how the Text Editor window underlines words not in its dictionary in the style of Microsoft Word. Suspect characters are highlighted yellow. Right-clicking on highlighted words presents a choice of corrections. It's also possible to edit and reformat text freely, as with a word processor. We found this much easier than grappling with a spellchecker dialog box, especially after we activated the Verifier option to display the relevant part of the scanned image for cross-reference. The advanced export options are bewildering; it took us a while to work out how to change the compression ratio of images in PDFs (the default setting produced small file sizes at the expense of image detail).

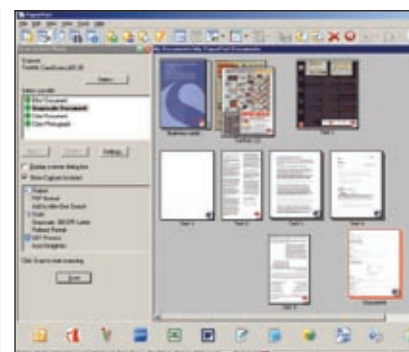
Recognition wasn't that quick, but was very accurate, losing ground to Abbyy FineReader only for 4-point text, where accuracy fell to 79 per cent. But OmniPage took the lead for faint-coloured text on white and black backgrounds. It handled unusual fonts well and gave the best comedy mistake, interpreting "In fact" as "Hi fart". Format preservation wasn't flawless, with some incorrect fonts for single words within a paragraph, but gave little cause for concern.

There's really little to choose between OmniPage and FineReader. However, OmniPage's superior manual correction facilities make it our favourite.

SUMMARY

- ✓ Excellent manual error correcting
- ✓ Easy to automate
- ✗ Struggles with very small text

OCR SOFTWARE Requires Windows 98 SE/Me/2000 (SP4)/XP/2003, Pentium processor, 128MB RAM (256MB recommended), 200MB disk space
PRODUCT CODE 2889X-W00-15.0
SUPPLIER www.pcvb.com
DETAILS www.nuance.co.uk



▲ Nuance PaperPort 11 offers basic and fully automatic OCR support

NUANCE Scansoft PaperPort 11

OVERALL RATING ★★★★★

OCR RATING ★★★

£34 inc VAT

PaperPort takes a holistic approach to paper management, with OCR, powerful search facilities, document browsing and previewing and the ability to annotate documents with virtual Post-it notes and highlighter pens. It appears as a printer driver to allow documents from other applications to be saved in PDF format. This potentially valuable feature is diluted because text appears as a bitmap rather than smooth fonts, even though editable text is hidden behind the image, so text looks scruffy.

Generating PDFs from scanned images is more versatile, with a choice of editable text behind or replacing the scanned image. But it's not possible to replace text areas with editable text while keeping images intact, nor to export OCR documents with images to Word. In fact, exporting to Word is poor, with only font formatting and no layout details preserved. Word documents are displayed in PaperPort without embedded images, but holding down Ctrl while double-clicking opens them in Word.

PaperPort's biggest flaw is its lack of proofing and spellchecking facilities. OCR is fully automatic and is performed as soon as the scan is complete. There's no way to intervene and correct mistakes. This isn't such a problem when exporting to a word processor as files can be edited after export, but for creating PDFs it's a serious hindrance. Character recognition was the fastest in our tests, but accuracy was the lowest when working with small font sizes.

PaperPort's basic OCR support is perhaps forgivable in light of its low price and range of other features. However, the poor results of its OCR output diminish the usefulness of its search and other facilities, too. **CS**

SUMMARY

- ✓ Powerful search and file-management features
- ✗ Basic OCR options
- ✗ Poor accuracy with small fonts

OCR SOFTWARE Requires Windows 2000 (SP4)/XP (SP2), Pentium processor, 128MB RAM (256MB recommended), 250MB disk space
PRODUCT CODE 6809X-W00-11.0
SUPPLIER www.misco.co.uk
DETAILS www.nuance.co.uk